

## ARTICLE OPEN



## Molecular Diagnostics

## Development and validation of a circulating microRNA panel for the early detection of breast cancer

Ruiyang Zou<sup>1</sup>, Sau Yeen Loke<sup>2</sup>, Yew Chung Tang<sup>1</sup>, Heng-Phon Too<sup>3,4</sup>, Lihan Zhou<sup>1</sup>✉, Ann S. G. Lee<sup>1,2,5,6</sup>✉ and Mikael Hartman<sup>7,8</sup>

© The Author(s) 2021

**BACKGROUND:** Mammography is widely used for breast cancer screening but suffers from a high false-positive rate. Here, we perform the largest comprehensive, multi-center study to date involving diverse ethnic groups, for the identification of circulating miRNAs for breast cancer screening.

**METHODS:** This study had a discovery phase ( $n = 289$ ) and two validation phases ( $n = 374$  and  $n = 379$ ). Quantitative PCR profiling of 324 miRNAs was performed on serum samples from breast cancer (all stages) and healthy subjects to identify miRNA biomarkers. Two-fold cross-validation was used for building and optimising breast cancer-associated miRNA panels. An optimal panel was validated in cohorts with Caucasian and Asian samples. Diagnostic ability was evaluated using area under the curve (AUC) analysis.

**RESULTS:** The study identified and validated 30 miRNAs dysregulated in breast cancer. An optimised eight-miRNA panel showed consistent performance in all cohorts and was successfully validated with AUC, accuracy, sensitivity, and specificity of 0.915, 82.3%, 72.2% and 91.5%, respectively. The prediction model detected breast cancer in both Caucasian and Asian populations with AUCs ranging from 0.880 to 0.973, including pre-malignant lesions (stage 0; AUC of 0.831) and early-stage (stages I–II) cancers (AUC of 0.916).

**CONCLUSIONS:** Our panel can potentially be used for breast cancer screening, in conjunction with mammography.

*British Journal of Cancer* (2022) 126:472–481; <https://doi.org/10.1038/s41416-021-01593-6>

**BACKGROUND**

Mammography has been widely-used as a screening tool for breast cancer despite its high false-positive rate, and its lack of sensitivity in detecting cancer in dense breasts [1]. A high rate of false positivity of 11–12% has been detected among women in the United States who have undergone mammographic screening [2, 3]. Minimally invasive methods such as miRNA-based liquid biopsies can potentially overcome these disadvantages and improve overall detection accuracy [4, 5]. MiRNAs are deemed suitable as biomarkers because of altered miRNA expression profiles in cancer that reflect disease development, as well as the stability and the accessibility of circulating miRNAs in a myriad of body fluids including blood, urine and saliva [6, 7].

MiRNAs are evolutionary conserved, single-stranded non-coding RNAs of 19–25 nucleotides that primarily function in mediating the degradation or translational repression of mRNA targets [7]. Under normal physiological conditions, miRNAs are key components of feedback mechanisms for a wide range of biological pathways such as cell proliferation, differentiation and apoptosis [8]. Conversely, dysregulated miRNAs have been

implicated in carcinogenesis including supporting tumour growth by inhibiting growth suppression, sustaining proliferative signalling and resisting cell death, activating invasion and metastasis, and promoting angiogenesis [6]. It is now known that miRNAs regulate oncogenesis through their tumour suppressor or oncogenic activities, with increasing evidence of aberrant miRNA expression in a variety of malignancies [9].

In an effort to improve breast cancer detection, numerous promising blood-derived miRNA biomarkers with superior discriminative ability as compared to mammography, have been reported in recent years [1, 10, 11]. The miRNAs miR-145, miR-21 and miR-221 are among the more frequently reported candidates and demonstrate potential for the early detection of breast cancer [12–15]. However, despite these findings, these breast cancer-associated miRNA biomarker studies are still at the discovery phase, with none successfully transitioned to clinical trials. This could be due to various shortcomings. For example, the majority of previously published circulating miRNA biomarker studies for breast cancer were conducted in smaller sample sizes comprising of a single ethnic group, and with one or no validation phase.

<sup>1</sup>Department of Research and Development, MiRXES Lab, Singapore, Singapore. <sup>2</sup>Cellular and Molecular Research, Humphrey Oei Institute of Cancer Research, National Cancer Centre, Singapore, Singapore. <sup>3</sup>Department of Biochemistry, Yong Loo Lin School of Medicine, National University of Singapore, Singapore, Singapore. <sup>4</sup>NUS Center for Cancer Research, Yong Loo Lin School of Medicine, National University of Singapore, Singapore, Singapore. <sup>5</sup>SingHealth Duke-NUS Oncology Academic Clinical Programme, Duke-NUS Medical School, Singapore, Singapore. <sup>6</sup>Department of Physiology, Yong Loo Lin School of Medicine, National University of Singapore, Singapore, Singapore. <sup>7</sup>Department of Surgery, Yong Loo Lin School of Medicine, National University of Singapore and National University Health System, Singapore, Singapore. <sup>8</sup>Saw Swee Hock School of Public Health, National University of Singapore and National University Health System, Singapore, Singapore. ✉email: lihan\_zhou@mirxes.com; dmslsg@nccs.com.sg

Received: 20 April 2021 Revised: 5 August 2021 Accepted: 6 October 2021

Published online: 10 January 2022

Several studies on blood-based circulating miRNA signatures were generated based on one discovery cohort, with the sample size of less than 200 study subjects without subsequent validation cohorts [11–14, 16]. Other studies had only one validation phase each [17–20]. A prospective validation using 116 women [21] was carried out as a follow-up to the study by Kodahl et al. [18]. However, in the prospective study, the nine-miRNA signature (miR-15a, miR-18a, miR-107, miR-133a, miR-139-5p, miR-143, miR-145, miR-365 and miR-425) failed to differentiate between the non-cancer controls and the ER+ or the ER– subtypes (AUC of 0.580 and 0.610, respectively) [21], possibly due to insufficient validation of the miRNA signature in the previous retrospective study.

Here, we have carried out the largest comprehensive, multi-center study to date involving diverse ethnic groups comprising of Caucasians and Asians, for the identification of circulating miRNAs associated with breast cancer. Expression profiles of serum miRNAs were elucidated from several independent cohorts of breast cancer cases and non-cancer controls following an analysis workflow that accounted for possible pre-analytic confounding factors. The performance of potential circulating miRNAs was established in three phases, with a discovery cohort and two independent validation cohorts. Hence, we are the first to report a miRNA-based breast cancer prediction model with high discriminative ability in classifying breast cancer patients from non-cancer individuals, and with the capability of identifying breast cancer in both Caucasian and Asian populations.

## METHODS

### Study design and cohorts

This multi-centre case-control study was carried out in three phases: one discovery phase and two validation phases. The three patient cohorts used (Discovery Cohort, Validation 1 Cohort, and Validation 2 Cohort, respectively) comprised samples from six different sources (Table 1). The Discovery cohort comprised European Caucasian samples obtained from the Asterand biobank (designated as Source 1) while the Validation 1 and 2 cohorts were a mix of Caucasian and Asian samples from five different sources in USA, Ukraine, Russia, and Singapore (designated as Sources 2–5) (Table 1). The Discovery and Validation cohorts included female subjects diagnosed with stage 1 to stage 3 breast cancer of all subtypes (Supplementary Table S1). The inclusion criteria for cases was women diagnosed with breast cancer, and the inclusion criteria for controls was no history of cancer in healthy female individuals. A blinded approach was not done as our study design included two validation phases. Written informed consent was obtained from all participants and the research was approved by all relevant Institutional Review Boards (IRBs). Samples from the Asterand and Tissue Solution biobanks were ethically collected under IRB-approved protocols and fully consented.

### Blood collection and serum processing

Peripheral blood samples (20 ml) were drawn from subjects using venipuncture and collected in serum tubes. Blood samples were clotted for 30–60 min and were centrifuged at 1300 rcf at room temperature for 20 min. Sera were then aliquoted for immediate storage at  $-80^{\circ}\text{C}$ .

### RNA Isolation

Total RNA was extracted from 200  $\mu\text{l}$  of each serum sample using the miRNeasy Serum/Plasma Kit (Qiagen, Hilden, Germany). This was done according to the manufacturer's recommendations, except for the following modifications: (a) a set of three proprietary spike-in controls (MiRiXES, Singapore) was added, representing high, medium, and low levels of RNA, into the sample lysis buffer (QIAzol Lysis Reagent, Qiagen) prior to sample RNA isolation. The spike-in controls are 20-nucleotide RNAs with unique sequences (distinct from any of the 2588 annotated mature human miRNAs in miRBase version 21.0, RRID:SCR\_003152) and are used to monitor RNA isolation efficiency and normalise for technical variations during RNA isolation; (b) bacteriophage MS2 RNA was added into sample lysis buffer (1  $\mu\text{g}$  per ml of QIAzol) to improve RNA isolation yield; (c) the samples were centrifuged at  $18,000 \times g$  for 15 min at room temperature after mixing with chloroform; and finally, (d) the RNA was eluted in 25  $\mu\text{l}$  of RNase-free water.

### RT-qPCR Detection of miRNA expression

For biomarker discovery, a highly-controlled RT-qPCR workflow was used to quantify the expression of 324 miRNAs in each serum sample (Supplementary Table S2). Serum RNA was reverse transcribed using miRNA-specific reverse transcription (RT) primers according to manufacturer's instructions (ID3EAL Customized Individual miRNA RT Primer, MiRiXES) on a Veriti™ Thermal Cycler (Applied Biosystems, Foster City, CA, USA). Multiplexed RT reactions were carried out using specific RT primers for 324 miRNAs. This proprietary list of 324 circulating miRNAs was selected based on experimental analysis of more than 1000 high confidence human miRNAs from several hundred serum and plasma specimens. These 324 miRNAs are therefore those which have been detected with high confidence in human serum and plasma samples. The RT primers were divided into 10 multi-plex primer pools (50–60-plex per pool) to minimise non-specific cross-overs and primer-primer interactions. For each RNA sample, 10 multiplex RT reactions were performed, each with 2  $\mu\text{l}$  of isolated RNA. Synthetic templates for standard curves of each miRNA (6-log serial dilution of 10 million to 100 copies) and a non-template control (nuclease-free water spiked with MS2) were reverse transcribed concurrently with the isolated sample RNA. Synthetic miRNA standard curves were used to absolutely quantify sample miRNA expression copy numbers. To measure 324 miRNAs using quantitative PCR (qPCR), all cDNAs, including those from synthetic miRNA standards, were pre-amplified using a 14-cycle PCR reaction with Augmentation Primer Pools (MiRiXES) on the Veriti™ Thermal Cycler. Single-plex qPCR was then performed on the amplified cDNA samples using a miRNA-specific qPCR assay (MiRiXES) and ID3EAL miRNA qPCR Master Mix according to manufacturer's instructions (MiRiXES). The qPCR reactions with technical duplicates were carried out on the ViiA™ qPCR system (384-well configuration, Applied Biosystems). Raw threshold cycle (Ct) values were calculated using the ViiA™ 7 RUO software with automatic baseline setting and a threshold of 0.5. RT-qPCR efficiency and potential cDNA amplification bias were assessed by analysing the Ct values of the synthetic miRNA standards. The absolute expression of each miRNA (number of copies present) in the serum sample was calculated by interpolation of sample Ct values with synthetic miRNA standard curves and correcting for variations in RT-qPCR efficiency. For biomarker validation, miRNA expression was quantified using the same workflow described above, adjusted for the number of miRNAs to be quantified.

### Biomarker discovery

The absolute quantities of 324 candidate miRNAs in the serum of both breast cancer cases and non-cancer controls were determined. The geNORM (geNORM, RRID:SCR\_006763) [22] and NormFinder (NormFinder, RRID:SCR\_003387) [23] software were used to identify endogenous reference miRNAs that had stable expression across all samples and could be used to normalise for varying sample RNA inputs for RT-qPCR. Three miRNAs with stable expression were identified and used to normalise the expression levels of miRNAs across samples: miR-128-3p, miR-652-3p, and miR-106b-3p (Supplementary Fig. S1). The normalised miRNA expression values were used to compare the expression levels of individual miRNAs between breast cancer cases and non-cancer controls. Unsupervised hierarchical clustering was carried out based on Euclidean distance of normalised miRNA expression levels in two dimensions (samples and miRNA expression). The top miRNAs with  $p < 0.01$  and magnitude of  $\log_2$  fold change  $> 0.5$  were selected for validation using the Validation 1 cohort. Statistical significance of differences in miRNA expression was determined using Student's *t*-test. All *p*-values were corrected for multiple hypotheses testing using false discovery rate (FDR) adjustment [24, 25]. Those miRNAs which were differentially expressed in both the Discovery and Validation 1 cohorts were considered validated. A relaxed cut-off of  $p < 0.05$  with magnitude of  $\log_2$  fold change  $> 0.5$  was used to identify validated miRNAs for biomarker panel building and optimisation.

### Biomarker panel building and optimisation

A two-fold cross-validation procedure that incorporated the sequential forward floating search (SFFS) algorithm [26] and a logistic regression model was used for building and optimising miRNA biomarker panels to discriminate between breast cancer cases and non-cancer controls. Starting from an empty panel, SFFS progressively searches for the next best marker to add to the panel already chosen. It is a search strategy to arrive at an optimal panel without conducting an exhaustive search of all possible combinations. For each iteration of the cross-validation procedure, SFFS will arrive at one optimal panel (one optimal two-miRNA panel, one optimal three-miRNA panel, and so on). The SFFS was used to select miRNA biomarkers for inclusion in each biomarker panel built. In each iteration of the two-fold cross validation procedure, the samples included

**Table 1.** Patient cohorts used in study.

Cohort:			Discovery		Validation 1		Validation 2	
Non-cancer (NC)/Cancer (C):			NC	C	NC	C	NC	C
			106	183	197	177	199	180
Source		Ethnicity						
1	Asterand biobank (Europe)	Caucasian	106 (100%)	183 (100%)	–	–	–	–
2	Asterand biobank (USA)	Caucasian	–	–	39 (19%)	39 (22%)	39 (19%)	40 (22%)
3	Tissue Solutions biobank (Ukraine)	Caucasian	–	–	33 (17%)	23 (13%)	34 (17%)	24 (13%)
4	Tissue Solutions biobank (Russia)	Caucasian	–	–	47 (24%)	48 (27%)	47 (24%)	48 (27%)
5	National University Hospital	Asian (Singapore)	–	–	35 (18%)	38 (21%)	35 (18%)	38 (21%)
6	National Cancer Centre Singapore	Asian (Singapore)	–	–	43 (22%)	29 (16%)	44 (22%)	30 (17%)
Age (years)								
Mean			53.7	52.4	53.0	56.6	55.5	55.4
Median			53	51	52	57	56	55
Range			42–65	30–85	29–82	31–77	26–83	28–87
Sex								
Male			0	0	0	0	0	0
Female			106	183	197	177	199	180
Cancer Stage								
0			–	0 (0%)	–	16 (9%)	–	23 (13%)
I			–	77 (42%)	–	51 (29%)	–	45 (25%)
II			–	78 (43%)	–	61 (34%)	–	58 (32%)
III			–	28 (15%)	–	5 (3%)	–	8 (4%)
IV			–	0 (0%)	–	3 (2%)	–	3 (2%)
Unknown			–	0 (0%)	–	41 (23%)	–	43 (24%)

in the combined Discovery and Validation 1 cohorts (comprising a total of 663 samples from six sources) were randomly partitioned into two equal groups: Group A and Group B. The proportion of subjects from each of the six sources were partitioned equally in both Group A and B. During each iteration of cross-validation, Group A was first used as the training set for building a breast cancer prediction model while Group B was used as the test set. The group assignments as training and testing sets were then swapped. For every multi-miRNA biomarker panel optimised in each iteration, a logistic regression prediction model was built and the diagnostic ability of each panel was evaluated using the area under the curve of the receiver operating characteristics (AUC) analysis. The cross-validation procedure was carried out 200 times. Thus, 200 two-miRNA panels, 200 three-miRNA panels, and so on, were optimised and tested. The diagnostic power (AUC) of each optimised multi-miRNA panel for classifying breast cancer and non-cancer patient samples was then calculated and compared with other panels optimised in each iteration. Using a logistic regression model incorporating multi-miRNA biomarker panel expression measurements, a prediction algorithm score could be calculated for each sample, with higher scores indicating increased risk of cancer. A prediction algorithm score cut-off was then used to predict breast cancer.

## RESULTS

### Discovery and validation of significant differentially regulated miRNAs

The expression levels of 324 miRNAs, which have been previously detected with high confidence in human serum, were quantified in the Discovery Cohort of 289 Caucasian samples (183 breast cancer samples and 106 non-cancer controls). All samples in this

cohort were obtained from a single source as shown in Table 1. MiRNAs that were significantly differentially expressed between breast cancer cases and non-cancer controls were identified by *p*-value of unpaired Student's *t*-test and fold change in expression. A total of 86 miRNAs that were differentially expressed between cancer cases and non-cancer controls with log<sub>2</sub> (fold change) more than 0.5 or less than –0.5 and *p*-value < 0.01 were identified (Fig. 1a).

The ability of these 86 differentially expressed miRNAs to differentiate between breast cancer cases and non-cancer controls was also assessed using AUC analysis. Out of these 86 differentially expressed miRNAs, 33 miRNAs had AUC > 0.5 and were selected for validation in a mixed Caucasian-Asian cohort (Validation 1). This cohort comprised of 374 samples (177 breast cancer cases and 197 non-cancer controls) from five different sources (three Caucasian and two Asian populations) as shown in Table 1.

Unsupervised hierarchical clustering based on the differential expressions of the 33 top-ranked miRNA biomarker candidates was carried out on the combined Discovery and Validation 1 cohort (663 samples comprising 360 breast cancer cases and 303 non-cancer controls). The cancer samples and the non-cancer samples were partially separated after clustering based on differential expression of these 33 miRNAs (Fig. 1b). Additionally, samples from the same sources were not clustered together based on their expression of these 33 miRNAs (Fig. 1b).

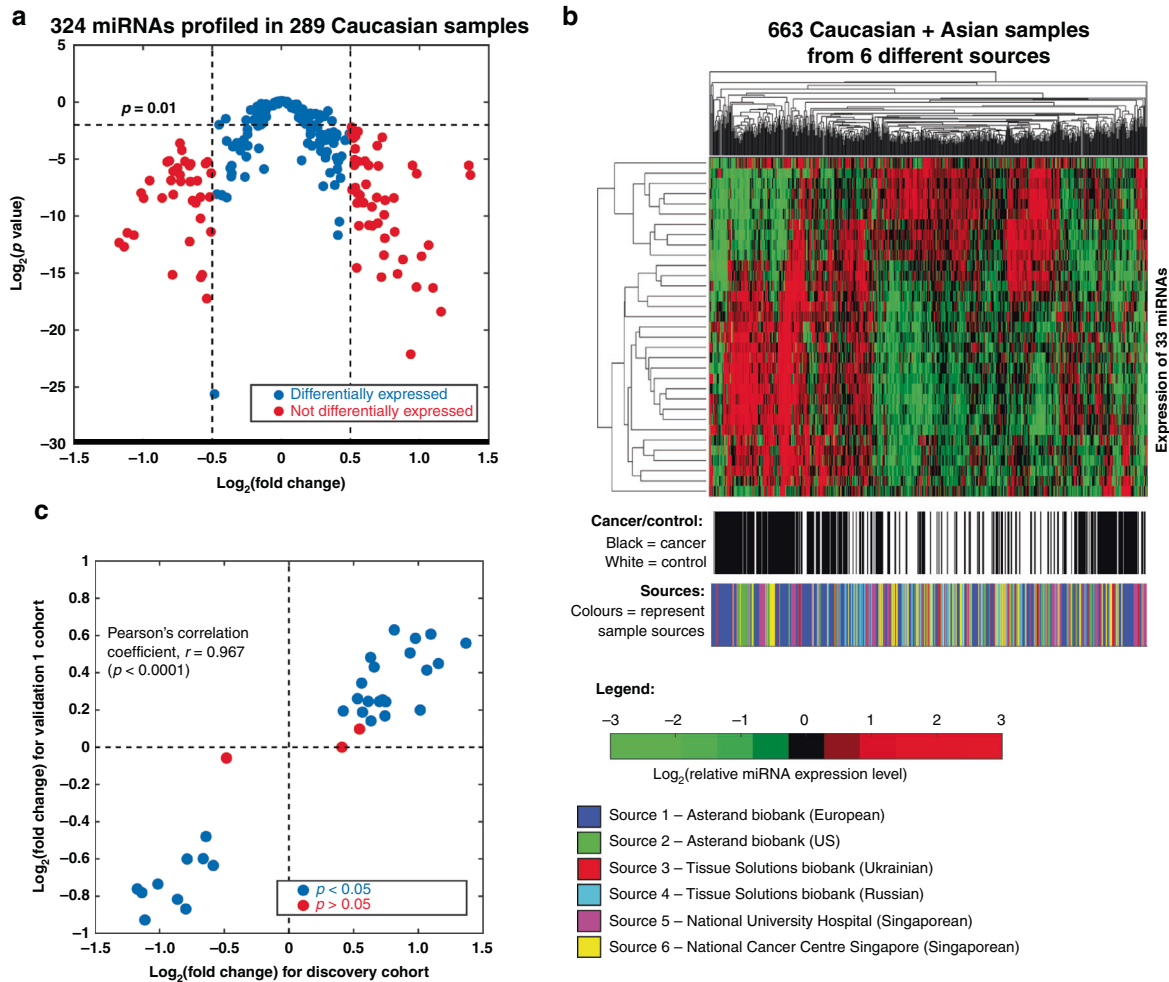
The log<sub>2</sub> (fold change) calculated for these 33 miRNA biomarker candidates in the Discovery cohort and Validation 1 cohort were compared and had a Pearson's correlation coefficient, *r* = 0.967

( $p < 0.0001$ ). Out of the 33 biomarker candidates identified from the Discovery cohort, 30 miRNAs were differentially expressed in breast cancer cases compared to non-cancer controls ( $p < 0.05$  by unpaired  $t$ -test) in the Validation 1 cohort (Fig. 1c). Three miRNA biomarker candidates that showed non-significant differential expressions ( $p > 0.05$ ) in Validation 1 cohort were excluded from the subsequent analysis, while the remaining 30 miRNA biomarker candidates were used for the biomarker panel optimisation phase.

### Optimisation of miRNA biomarker panels

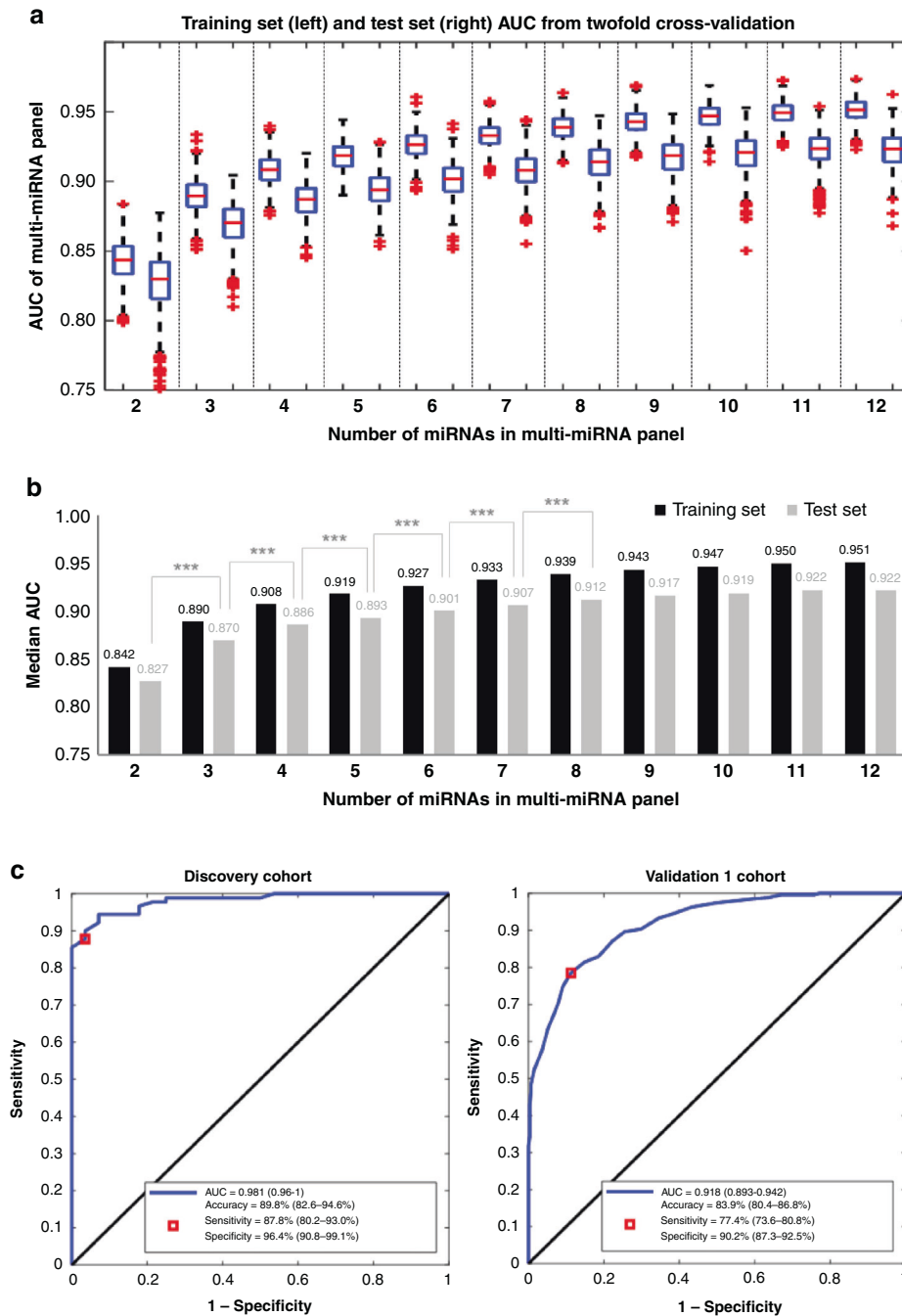
To identify an optimal panel with good performance while balancing the number of miRNAs included for the practicality of clinical testing, multi-miRNA panels were assessed. The best-performing multi-miRNA panel comprising between two to twelve miRNAs were formed from the 30 validated miRNA biomarker candidates using a two-fold cross-validation procedure that incorporated a feature selection algorithm (SFSS). AUC of miRNA panel performance in the training and test group was calculated for 200 iterations of cross-validation with multi-miRNA panels comprising two to twelve miRNAs (Fig. 2a). The median AUC for breast cancer prediction from 200 iterations of training and testing

was calculated for each set of cross-validation experiments comprising two to twelve miRNAs (Fig. 2b). The median AUC increased significantly ( $p < 0.001$ ) with increasing number of miRNAs in the biomarker panels that consisted of two to eight miRNAs, until it reached a plateau after the inclusion of eight miRNAs on the panel (Fig. 2b); hence indicating that eight miRNAs is the optimal number of biomarkers to be included on the panel. The addition of more miRNAs did not lead to a statistically significant increase in AUC. The optimal eight-miRNA biomarker panel with the highest AUCs of 0.981 and 0.918 in the Discovery and Validation 1 cohorts, respectively, was chosen for further validation (Fig. 2c). This optimal panel included miR-133a-3p, miR-497-5p, miR-24-3p, and miR-125b-5p, which were upregulated in breast cancer cases compared to controls, and miR-377-3p, miR-374c-5p, miR-324-5p and miR-19b-3p, which were downregulated in breast cancer cases as compared to controls (Supplementary Table S3). At the point of maximum classification accuracy, sensitivity and specificity were 87.8% (95% CI, 80.2–93.0%) and 96.4% (95% CI, 90.8–99.1%) in the Discovery Cohort, and 77.4% (95% CI, 73.6–80.8%) and 90.2% (95% CI, 87.3–92.5%) in the Validation 1 Cohort (Fig. 2c).



**Fig. 1 Biomarker discovery.** **a** Volcano plot for 324 miRNAs profiled in 183 breast cancer patients as compared to 106 healthy individuals in the discovery cohort. Eighty-six miRNAs with  $p$ -values of less than 0.01 and magnitudes of  $\text{log}_2(\text{fold change})$  of more than 0.5 are highlighted in red. **b** Heat-map of 663 cancer and non-cancer samples clustered using the expression of 33 selected miRNA biomarkers identified in the Discovery cohort. The expression levels (copy/ml) of miRNAs are presented in  $\text{log}_2$  scale and standardised to zero mean. The colour scale represents the concentrations of miRNA. Hierarchical clustering was carried out for both dimensions (miRNAs: Y-axis, samples: X-axis) based on the Euclidean distance. **c** Correlation of  $\text{log}_2(\text{fold change})$  for 33 selected miRNA biomarker candidates in the Discovery cohort and Validation 1 cohort. The miRNA biomarker candidates with  $p$ -values larger than 0.05 in the validation 1 cohort are highlighted in red.





**Fig. 2 Optimisation of multi-miRNA biomarker panels.** **a** Boxplots of AUC of multi-miRNA biomarker panels (with 2–12 miRNAs), in both the training and test sets, calculated from 200 iterations of the two-fold cross-validation process. The boxplot presents the 25th, 50th, and 75th percentiles in panel AUC. **b** Median AUC for the training and test sets from 200 iterations of the two-fold cross-validation process for multi-miRNA panels with 2–12 miRNAs.  $***p < 0.001$  (Student's *t*-test). **c** ROC curves for breast cancer prediction performance of the optimal eight-miRNA biomarker panel in the Discovery and Validation 1 cohorts. The point with the maximum classification accuracy is shown as the red box. The sensitivity and specificity values at the maximum accuracy point are also shown. The 95% CI for these values is shown in the brackets.

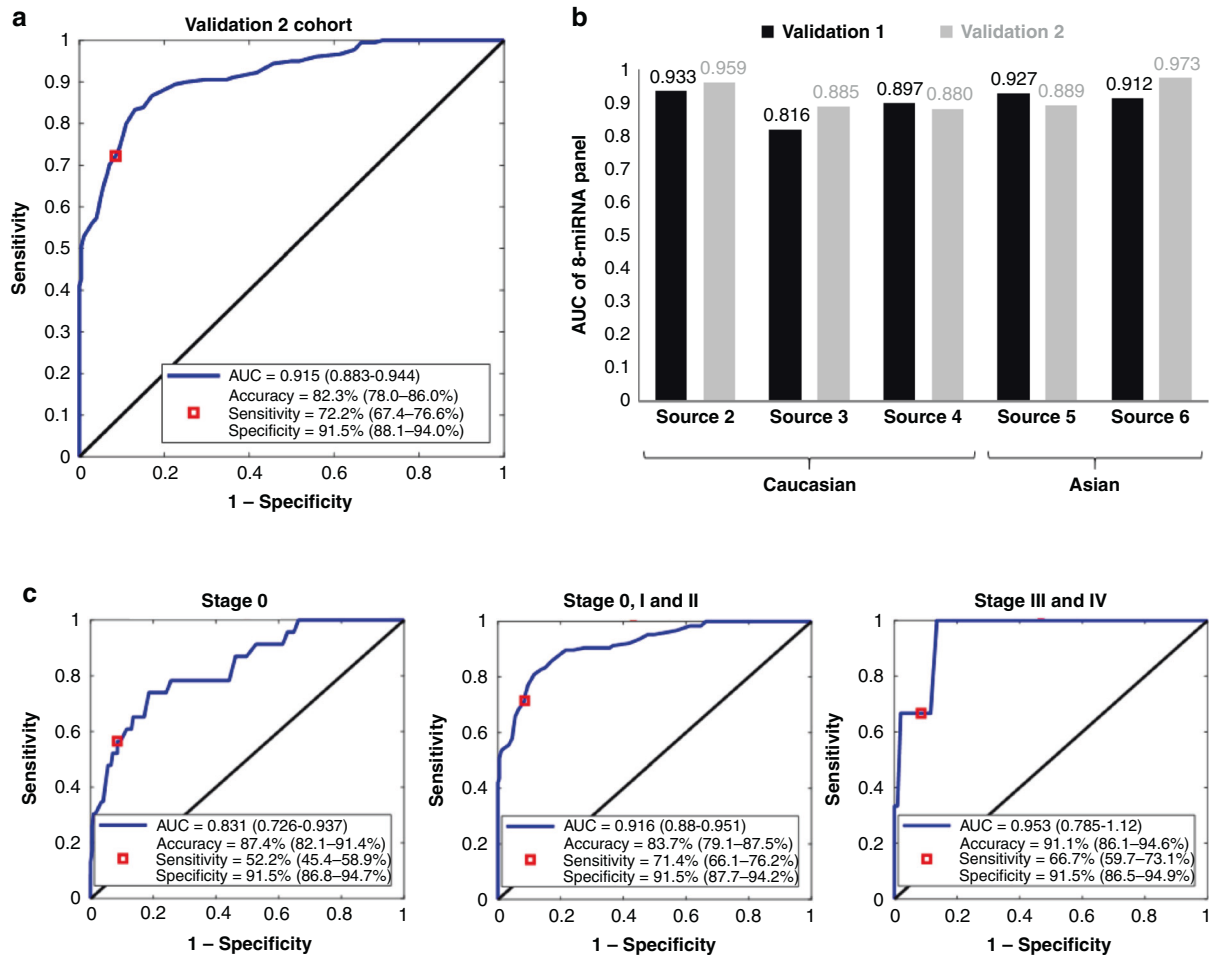
### Validation of optimal eight-miRNA biomarker panel signature

Validation of the optimised eight-miRNA biomarker panel signature was carried out in the Validation 2 cohort which comprised of 379 samples (180 breast cancer and 199 non-cancer samples). The AUC of the eight-miRNA biomarker panel in classifying breast cancer and non-cancer samples was 0.915 (95% CI, 0.883–0.944) (Fig. 3a). At the point of maximum classification accuracy, sensitivity was 72.2% (95% CI, 67.4–76.6%) with a specificity of 91.5% (95% CI, 88.1–94.0%). When the Validation 1 and 2 cohorts were separated by sample source, the eight-miRNA biomarker

panel had AUC ranging from 0.816 to 0.973 (Fig. 3b). Performance was comparable between Caucasian and Asian sample sources (Fig. 3b) and for early stage (stages 0, I and II) and late stage (stages III and IV) breast cancers (Fig. 3c).

### Breast cancer prediction algorithm based on miRNA biomarker signature

A prediction algorithm based on a logistic regression model that takes into account the expression levels of the eight miRNAs in the biomarker panel was developed to calculate a cancer risk score.



**Fig. 3** Validation of optimal eight-miRNA biomarker panel. **a** ROC curve for breast cancer prediction performance of the optimal eight-miRNA biomarker panel in the Validation 2 cohort. The point with the maximum classification accuracy is shown as the red box. The sensitivity and specificity values at the maximum accuracy point are also shown. The 95% CI for these values is shown in the brackets. **b** AUC of eight-miRNA biomarker panel in detecting breast cancer in the Validation 1 and Validation 2 cohorts separated by sample source. **c** ROC curves for performance of the optimal eight-miRNA biomarker panel in predicting early (stages 0, I and II) and late (stages III and IV) breast cancer in the Validation 2 cohort.

The median cancer risks score calculated from the prediction algorithm were higher for cancer samples compared to non-cancer samples across all cohorts regardless of sample source (Fig. 4a). The panel effectively detects breast cancer of all stages, including early stage breast cancers (stages 0, I and II) (Fig. 3c), with cancer risk scores from breast cancer samples of all stages falling in the same range that is higher than that of non-cancer samples (Fig. 4b). The distribution of breast cancer samples by stage in each cohort is shown in Table 1.

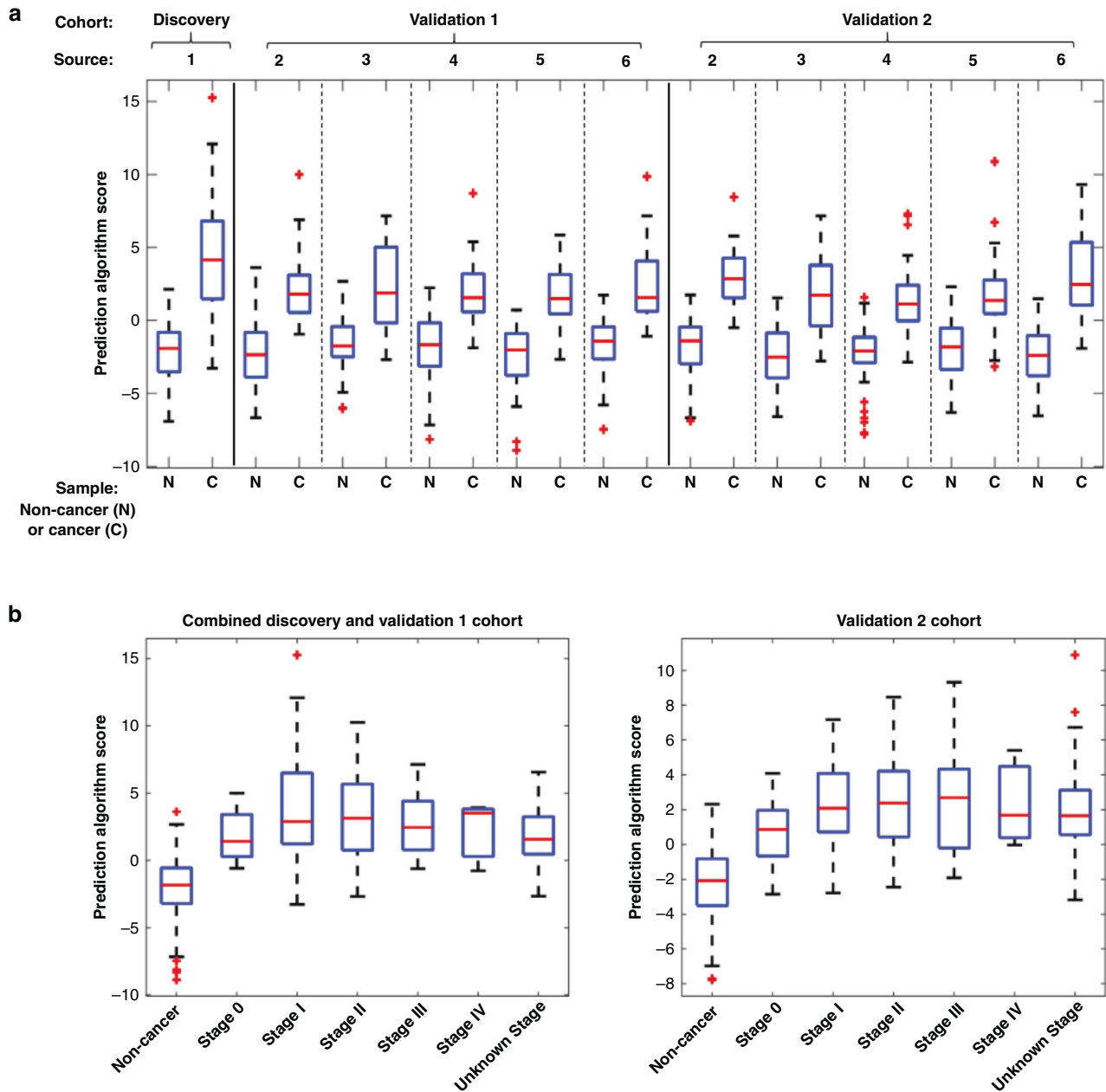
## DISCUSSION

We describe here an eight-miRNA biomarker signature that can differentiate between breast cancer patients and non-cancer individuals, which was derived from a multi-ethnic study comprising of a discovery phase followed by two validation phases. Our miRNA-based signature could perform well in distinguishing breast cancer cases from non-cancer controls as evidenced by a high AUC of 0.915 achieved in validation phase 2, which comprised of both Caucasians and Asians. These results imply that the performance of our model is robust, accurate and effective in classifying individuals diagnosed with breast cancer from those who are cancer-free. Therefore, to our knowledge, the present study is the first to date to

describe the largest comprehensive multicenter study for the development of a circulating miRNA-based breast cancer prediction model that is applicable to both Caucasian and Asian populations.

The performance of the present miRNA-based prediction model for the initial discovery and two validation phases were consistent, as demonstrated by their respective AUCs of 0.981, 0.918 and 0.915. Likewise, when the validation phases 1 and 2 were analysed based on the sub-cohorts obtained from different sample sources, the range of AUCs generated in these sub-cohorts for both phases were comparable, ranging from 0.816 to 0.933 in phase 1 and from 0.880 to 0.973 in phase 2. These results highlight the high reproducibility and accuracy of our model in differentiating breast cancer cases from non-cancer controls for both Caucasians and Asians, suggesting its potential universal usability for various ethnicities. Our current findings are an important advancement for breast cancer miRNA biomarker research as our study design, analysis and results are unprecedented in the field.

Among the eight miRNAs in our signature, miR-133a-3p, miR-497-5p, miR-24-3p, and miR-125b-5p were found to be upregulated, whereas miR-377-3p, miR-374c-5p, miR-324-5p and miR-19b-3p were found to be downregulated in breast cancer cases as compared to controls. All of these miRNAs, except miR-133a-3p, have been reported to be implicated in the pathophysiology of



**Fig. 4 Prediction algorithm based on the expression of eight-miRNA biomarker panel. a** Prediction algorithm scores of cancer and non-cancer samples calculated from the expression of eight-miRNA panel expression. **b** Prediction algorithm scores of non-cancer samples and cancer samples by tumour stage (0, I, II, III, IV, and unknown).

breast cancer. For example, both miR-24-3p and miR-125b-5p have been identified as potential breast cancer biomarkers for the early detection [27, 28], prognosis [29, 30], or prediction of recurrence [31, 32]. Among the eight miRNAs discovered, there are discrepancies reported between the present and previous studies regarding the expression levels of miR-497-5p in breast cancer. Based on our current observation, miR-497-5p was upregulated in the serum samples of breast cancer patients whereas several studies have reported the decreased expression of miR-497-5p in breast cancer tissue samples and cell lines [33–36]. In a nude mouse xenograft tumour model, the inhibitory role of miR-497-5p in tumour growth and angiogenesis has been demonstrated [33] while low miR-497-5p expression was associated with poor prognosis of breast cancer patients [34]. For miR-377-3p, studies have shown that miR-377-3p was one of the miRNA transcripts

that could predict tumour progesterone status with 100% accuracy [37] and the Linc00339/miR-377-3p/HOXC6 axis represented a novel pathway in the progression of triple-negative breast cancer [38]. MiR-374-5p has been shown to repress development of breast cancer through TATA-box binding protein associated factor 7 (TAF7)-mediated transcriptional regulation of DEP domain containing 1 (DEPDC1) [39]. The expression of miR-374-5p was downregulated in various breast cancer cell lines [39], similar to our observation in this study. A six-miRNA signature which included miR-324-5p has been discovered to be significantly associated with the reduced overall survival of triple-negative breast cancer [40]. MiR-19b-3p has also been shown to be downregulated in hormone receptor-positive/HER2-negative breast cancer [41]. With its high sensitivity and specificity in identifying breast cancer from healthy tissues and its the

involvement in regulation of genes in oncogenic pathways, miR-19b-3p may serve as a diagnostic marker or therapeutic target for breast cancer [41].

Although our model could not outperform the five-miRNA signature (miR-1246, miR-1307-3p, miR-4634, miR-6861-5p and miR-6875-5p) model reported by the largest breast cancer miRNA biomarker study to date with an AUC of 0.971, it is noted that the latter was based primarily on microarray profiling, a method which is known to have poor specificity [42], and that only one miRNA, miR-1246, was validated by qRT-PCR using 26 serum samples [10]. For the performance comparison of the prediction model in a Singaporean Chinese cohort, our present model was able to achieve a better classification (AUC of 0.973) than the two-miRNA combinations generated from miR-1, miR-92a, miR-133a and miR-133b that were previously published (AUC of 0.900–0.910) [43]. In another study, a plasma signature of miR-145 and miR-451 was identified and demonstrated an outstanding ability in distinguishing breast cancer patients from healthy controls (AUC of 0.931) following a blind validation [15]. Although Ng et al. [15] conducted the study using multiple cohorts in different phases [17], the total sample size was smaller than the present study. In addition, a recent study reported a six-miRNA signature (miR-3124-5p, miR-1184, miR-4423-3p, miR-4529-3p, miR-7855-5p, and miR-4446-3p) with an AUC of 0.896 for discriminating high-risk women who eventually diagnosed with breast cancer from those who remained cancer-free [44]. However, a downside of this study was that the model was only tested in a single cohort of 48 samples [44].

Another observation from these studies is that there is a lack of strong overlap of miRNAs between studies which could possibly be attributed to differences between studies in sample type (whole blood, plasma or serum), timing of blood collection (before or after surgery), technology platform (microarray, RT-PCR or next-generation sequencing), study design and differences in data analysis. Hence, this suggests that for biomarker discovery research, having multiple validation cohorts is of utmost importance in order to verify the biomarker signature.

This study utilised qRT-PCR for miRNA profiling, since qRT-PCR is deemed as the gold standard for nucleic acid quantification due to the sensitivity and specificity of the method [42, 45]. Standard curves were also used to determine the absolute expression (copy number) of each miRNA target. In addition, since qRT-PCR is commonly utilised in various multigene prognostic assays including Oncotype DX [46], Breast Cancer Index [47], and EndoPredict [48], this makes our miRNA-based breast cancer prediction model more readily translatable as a molecular diagnostic assay for clinical use.

Apart from miRNA biomarkers, there are other notable research efforts assessing alternative blood-based bioanalytes for breast cancer detection, such as the CancerSEEK study [49] and the Circulating Cell-Free Genome Atlas (CCGA) study [50]. CancerSEEK is a pan-cancer blood test intended for the identification of eight cancer types including breast cancer, by evaluating mutations in 16 genes from cell-free DNA (cfDNA) and the expression of eight protein biomarkers using multiplex PCR followed by next-generation sequencing and immunoassays respectively [49]. Similarly, the CCGA study which is an on-going prospective longitudinal cohort study that has enrolled approximately 15,000 study participants, also aims to develop a multi-cancer detection blood test by profiling cfDNA using sequencing-based methods [50]. Although these assays have been tested to detect different cancer types and stages, their performance for identifying breast cancer, especially in the early stages, is still under par. For the CancerSEEK test, the median detection sensitivity for breast cancer was 33% as compared to 98% for ovarian cancer, whereas the median detection sensitivity for stage I of all cancer types was only 43% as compared to 78% for stage III cancers [49]. Moreover, the tests developed by the CCGA study were poor in

identifying various breast cancer molecular subtypes with sensitivities below 60% [51]. In contrast, our miRNA-based model showed superior discrimination performance, even for differentiating between healthy controls and those at pre-malignant stages (stage 0) with the AUC, accuracy, sensitivity and specificity of 0.831, 87.4%, 52.2% and 91.5%, respectively. In addition, the AUC and sensitivity increased to 0.916 and 71.4%, respectively for the detection of the pre-malignant stage and early-stage breast cancers (stages 0–II).

There are limitations or challenges in this current study. Firstly, while the present circulating miRNA-based prediction model performed well in identifying breast cancer cases, the specificity of the model for breast cancer as compared to other cancer types warrants further investigation [1]. Secondly, pre-analytical confounding factors might be present in the study given that these samples were sourced from different clinical centers and some pre-analytical factors such as sample processing procedure, storage duration, and transport condition could differ across the centers [1]. Hence, for future studies, a standardised protocol for sample collection, transport, processing, and storage should be used to minimise possible confounding factors that may lead to experimental biasness in the analysis of the predictive performance of the model. Thirdly, as the present study was conducted retrospectively, it is essential to further validate the miRNA-based prediction model in a prospective or blinded study and to benchmark the accuracy, sensitivity and specificity of our model against the current gold standard for breast cancer screening [1]. Such studies could be done on individuals who participate in routine annual or biennial mammogram screening, to ascertain its predictive value through comparison with mammography in a clinical setting. Future studies could also assess the performance of this panel in detecting breast cancers belonging to various subtypes, such as triple-negative breast cancers.

Together, this present study has generated a robust prediction model for breast cancer, applicable for Caucasian and Asian populations and patients of various cancer stages. The present miRNA-based prediction model can be potentially developed as an alternative modality for breast cancer screening, and may reduce the number of biopsies resulting from false-positive mammograms.

## DATA AVAILABILITY

The datasets used and analysed in this study are available from the corresponding authors on reasonable request.

## REFERENCES

1. Loke SY, Lee ASG. The future of blood-based biomarkers for the early detection of breast cancer. *Eur J Cancer*. 2018;92:54–68.
2. Ong MS, Mandl KD. National expenditure for false-positive mammograms and breast cancer overdiagnoses estimated at \$4 billion a year. *Health Aff (Millwood)*. 2015;34:576–83.
3. Nelson HD, O'Meara ES, Kerlikowske K, Balch S, Miglioretti D. Factors associated with rates of false-positive and false-negative results from digital mammography screening: An analysis of registry data. *Ann Intern Med*. 2016;164:226–35.
4. Ying L, Du L, Zou R, Shi L, Zhang N, Jin J, et al. Development of a serum miRNA panel for detection of early stage non-small cell lung cancer. *Proc Natl Acad Sci USA*. 2020;117:25036–42.
5. So, JBY, Kapoor, R, Zhu, F, Koh, C, Zhou, L, Zou, R, et al. Development and validation of a serum microRNA biomarker panel for detecting gastric cancer in a high-risk population. *Gut*. 2020; <https://doi.org/10.1136/gutjnl-2020-322065>.
6. Hayes J, Peruzzi PP, Lawler S. MicroRNAs in cancer: biomarkers, functions and therapy. *Trends Mol Med*. 2014;20:460–9.
7. Schwarzenbach H, Nishida N, Calin GA, Pantel K. Clinical relevance of circulating cell-free microRNAs in cancer. *Nat Rev Clin Oncol*. 2014;11:145–56.
8. Esquela-Kerscher A, Slack FJ. Oncomirs—microRNAs with a role in cancer. *Nat Rev Cancer*. 2006;6:259–69.
9. Kent OA, Mendell JT. A small piece in the cancer puzzle: microRNAs as tumor suppressors and oncogenes. *Oncogene*. 2006;25:6188–96.



10. Shimomura A, Shiino S, Kawauchi J, Takizawa S, Sakamoto H, Matsuzaki J, et al. Novel combination of serum microRNA for detecting breast cancer in the early stage. *Cancer Sci*. 2016;107:326–34.
11. Mar-Aguilar F, Mendoza-Ramirez JA, Malagon-Santiago I, Espino-Silva PK, Santuario-Facio SK, Ruiz-Flores P, et al. Serum circulating microRNA profiling for identification of potential breast cancer biomarkers. *Dis Markers*. 2013;34:163–9.
12. Thakur S, Grover RK, Gupta S, Yadav AK, Das BC. Identification of specific miRNA signature in paired sera and tissue samples of Indian women with triple negative breast cancer. *PLoS ONE*. 2016;11:e0158946.
13. Motawi TM, Sadik NA, Shaker OG, El Masry MR, Mohareb F. Study of microRNAs-21/221 as potential breast cancer biomarkers in Egyptian women. *Gene*. 2016;590:210–9.
14. Hannafon BN, Trigo YD, Calloway CL, Zhao YD, Lum DH, Welm AL, et al. Plasma exosome microRNAs are indicative of breast cancer. *Breast Cancer Res*. 2016;18:90.
15. Ng EK, Li R, Shin VY, Jin HC, Leung CP, Ma ES, et al. Circulating microRNAs as specific biomarkers for breast cancer detection. *PLoS ONE*. 2013;8:e53141.
16. Schwarzenbach H, Milde-Langosch K, Steinbach B, Muller V, Pantel K. Diagnostic potential of PTEN-targeting miR-214 in the blood of breast cancer patients. *Breast Cancer Res Treat*. 2012;134:933–41.
17. Cuk K, Zucknick M, Madhavan D, Schott S, Golatta M, Heil J, et al. Plasma microRNA panel for minimally invasive detection of breast cancer. *PLoS ONE*. 2013;8:e76729.
18. Kodahl AR, Lyng MB, Binder H, Cold S, Gravgaard K, Knoop AS, et al. Novel circulating microRNA signature as a potential non-invasive multi-marker test in ER-positive early-stage breast cancer: a case control study. *Mol Oncol*. 2014;8:874–83.
19. Zhang L, Xu Y, Jin X, Wang Z, Wu Y, Zhao D, et al. A circulating miRNA signature as a diagnostic biomarker for non-invasive early detection of breast cancer. *Breast Cancer Res Treat*. 2015;154:423–34.
20. Freres P, Wenric S, Boukerroucha M, Fasquelle C, Thiry J, Bovy N, et al. Circulating microRNA-based screening tool for breast cancer. *Oncotarget*. 2016;7:5416–28.
21. Lyng MB, Kodahl AR, Binder H, Ditzel HJ. Prospective validation of a blood-based 9-miRNA profile for early detection of breast cancer in a cohort of women examined by clinical mammography. *Mol Oncol*. 2016;10:1621–6.
22. Vandesompele J, De Preter K, Pattyn F, Poppe B, Van Roy N, De Paeppe A, et al. Accurate normalization of real-time quantitative RT-PCR data by geometric averaging of multiple internal control genes. *Genome Biol*. 2002;3:Research0034.
23. Andersen CL, Jensen JL, Orntoft TF. Normalization of real-time quantitative reverse transcription-PCR data: a model-based variance estimation approach to identify genes suited for normalization, applied to bladder and colon cancer data sets. *Cancer Res*. 2004;64:5245–50.
24. Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J R Stat Soc Ser B Stat Methodol*. 1995;57:289–300.
25. Benjamini Y, Yekutieli D. The control of the false discovery rate in multiple testing under dependency. *Ann Stat*. 2001;29:1165–88.
26. Xiong M, Fang X, Zhao J. Biomarker identification by feature wrappers. *Genome Res*. 2001;11:1878–87.
27. An X, Quan H, Lv J, Meng L, Wang C, Yu Z, et al. Serum microRNA as potential biomarker to detect breast atypical hyperplasia and early-stage breast cancer. *Future Oncol*. 2018;14:3145–61.
28. Matamala N, Vargas MT, Gonzalez-Campora R, Minambres R, Arias JI, Menendez P, et al. Tumor microRNA expression profiling identifies circulating microRNAs for early breast cancer detection. *Clin Chem*. 2015;61:1098–106.
29. Khodadadi-Jamayran A, Akgol-Oksuz B, Afanasyeva Y, Heguy A, Thompson M, Ray K, et al. Prognostic role of elevated mir-24-3p in breast cancer and its association with the metastatic process. *Oncotarget*. 2018;9:12868–78.
30. Inoronato M, Grimaldi AM, Mirabelli P, Cavaliere C, Parente CA, Franzese M, et al. Circulating miRNAs in Untreated Breast Cancer: An Exploratory Multimodality Morpho-Functional Study. *Cancers (Basel)*. 2019;11:876.
31. Kim C, Go EJ, Kim A. Recurrence prediction using microRNA expression in hormone receptor positive breast cancer during tamoxifen treatment. *Biomarkers*. 2018;23:804–11.
32. Bašová P, Pešta M, Sochor M, Stopka T. Prediction potential of serum miR-155 and miR-24 for relapsing early breast cancer. *Int J Mol Sci*. 2017;18:2116.
33. Wu Z, Cai X, Huang C, Xu J, Liu A. miR-497 suppresses angiogenesis in breast carcinoma by targeting HIF-1 $\alpha$ . *Oncol Rep*. 2016;35:1696–702.
34. Zhong H, Yang J, Zhang B, Wang X, Pei L, Zhang L, et al. LncRNA GACAT3 predicts poor prognosis and promotes cell proliferation in breast cancer through regulation of miR-497/CCND2. *Cancer Biomark*. 2018;22:787–97.
35. Li D, Zhao Y, Liu C, Chen X, Qi Y, Jiang Y, et al. Analysis of MiR-195 and MiR-497 expression, regulation and role in breast cancer. *Clin Cancer Res*. 2011;17:1722–30.
36. Li X, Wang Q, Rui Y, Zhang C, Wang W, Gu J, et al. HOXC13-AS promotes breast cancer cell growth through regulating miR-497-5p/PTEN axis. *J Cell Physiol*. 2019;234:22343–51.
37. Lowery AJ, Miller N, Devaney A, McNeill RE, Davoren PA, Lemetre C, et al. MicroRNA signatures predict oestrogen receptor, progesterone receptor and HER2/neu receptor status in breast cancer. *Breast Cancer Res*. 2009;11:R27.
38. Wang X, Chen T, Zhang Y, Zhang N, Li C, Li Y, et al. Long noncoding RNA Linc00339 promotes triple-negative breast cancer progression through miR-377-3p/HOXC6 signaling pathway. *J Cell Physiol*. 2019;234:13303–17.
39. Hao S, Tian W, Chen Y, Wang L, Jiang Y, Gao B, et al. MicroRNA-374c-5p inhibits the development of breast cancer through TATA-box binding protein associated factor 7-mediated transcriptional regulation of DEP domain containing 1. *J Cell Biochem*. 2019;120:15360–8.
40. Turashvili G, Lightbody ED, Tyryshkin K, SenGupta SK, Elliott BE, Madarnas Y, et al. Novel prognostic and predictive microRNA targets for triple-negative breast cancer. *FASEB J*. 2018; fj201800120R, <https://doi.org/10.1096/fj.201800120R>.
41. Maleki E, Ghaedi K, Shahanipour K, Karimi Kurdistani Z. Down-regulation of microRNA-19b in hormone receptor-positive/HER2-negative breast cancer. *APMIS*. 2018;126:303–8.
42. Pritchard CC, Cheng HH, Tewari M. MicroRNA profiling: approaches and considerations. *Nat Rev Genet*. 2012;13:358–69.
43. Chan M, Liaw CS, Ji SM, Tan HH, Wong CY, Thike AA, et al. Identification of circulating microRNA signatures for breast cancer detection. *Clin Cancer Res*. 2013;19:4477–87.
44. Farina NH, Ramsey JE, Cuke ME, Ahern TP, Shirley DJ, Stein JL, et al. Development of a predictive miRNA signature for breast cancer risk among high-risk women. *Oncotarget*. 2017;8:112170–83.
45. Schmittgen TD, Lee EJ, Jiang J, Sarkar A, Yang L, Elton TS, et al. Real-time PCR quantification of precursor and mature microRNA. *Methods*. 2008;44:31–38.
46. Knezevic D, Goddard AD, Natraj N, Cherbavaz DB, Clark-Langone KM, Snable J, et al. Analytical validation of the Oncotype DX prostate cancer assay - a clinical RT-PCR assay optimized for prostate needle biopsies. *BMC Genomics*. 2013;14:690.
47. Zhang Y, Schnabel CA, Schroeder BE, Jerevall PL, Jankowitz RC, Fornander T, et al. Breast cancer index identifies early-stage estrogen receptor-positive breast cancer patients at risk for early- and late-distant recurrence. *Clin Cancer Res*. 2013;19:4196–205.
48. Kronenwett R, Bohmann K, Prinzler J, Sinn BV, Haufe F, Roth C, et al. Decentral gene expression analysis: analytical validation of the Endopredict genomic multianalyte breast cancer prognosis test. *BMC Cancer*. 2012;12:456.
49. Cohen JD, Li L, Wang Y, Thoburn C, Afsari B, Danilova L, et al. Detection and localization of surgically resectable cancers with a multi-analyte blood test. *Sci (N. Y. N. Y.)*. 2018;359:926–30.
50. GRAIL. I Circulating Cell-free Genome Atlas Study, <https://grail.com/clinical-studies/circulating-cell-free-genome-atlas-study/> Accessed April 18, (2018).
51. Liu MC, Maddala T, Aravanis A, Hubbell E, Beausang JF, Filippova D, et al. Breast cancer cell-free DNA (cfDNA) profiles reflect underlying tumor biology: the circulating cell-free genome atlas (CCGA) study. *J Clin Oncol*. 2018;36:536–536.

## AUTHOR CONTRIBUTIONS

RZ conceptualised the study, designed and carried out experiments, analysed data and edited the manuscript. SYL curated data and helped write the manuscript. YCT analysed data and helped write the manuscript. HPT conceptualised the study and provided resources for experiments. LZ conceptualised the study, designed and carried out experiments, and provided resources for experiments. ASGL conceptualised the study, supervised the research, acquired funding, provided project administration, and helped write the manuscript. MH conceptualised the study, supervised the research, and provided resources for experiments.

## FUNDING INFORMATION

ASGL was supported in part by grants from the National Medical Research Council of Singapore (NMRC/EDG/1025/2010 and NMRC/CBRG/0087/2015). MH acknowledges support from the National University Cancer Institute Singapore (NCIS) Centre Grant [NMRC/CG/NCIS /2010 and NMRC/CG/012/2013] and a National University of Singapore start-up grant.

## COMPETING INTERESTS

HPT, LZ and RZ are shareholders of MiRXES Pte Ltd. LZ, RZ and YCT are employees of MiRXES Pte Ltd. All other authors declare no competing interests.

## ETHICS APPROVAL AND CONSENT TO PARTICIPATE

The study was conducted according to the guidelines of the Declaration of Helsinki, and approved by the Centralised Institutional Review Board (CIRB) of SingHealth (CIRB Ref: 2014/311/B; date of approval on 6 May 2014) and the Domain Specific Review Board (DSRB) of the National Healthcare Group (NHG) (NHG DSRB Ref: 2015/01118; date of approval on 18 January 2016).

**CONSENT TO PUBLISH**

Not applicable.

**ADDITIONAL INFORMATION**

**Supplementary information** The online version contains supplementary material available at <https://doi.org/10.1038/s41416-021-01593-6>.

**Correspondence** and requests for materials should be addressed to Lihan Zhou or Ann S. G. Lee.

**Reprints and permission information** is available at <http://www.nature.com/reprints>

**Publisher's note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



**Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2021